Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio

Francesca Bonin, Felice Dell'Orletta, Simonetta Montemagni e Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa

1. Introduzione

Il presente contributo descrive un approccio basato su strumenti di Trattamento Automatico del Linguaggio (TAL) finalizzato all'acquisizione di terminologia specialistica a partire da corpora di testi di dominio. La metodologia qui presentata prospetta una possibile soluzione alla principale difficoltà nel processo di estrazione terminologica automatica, ovvero l'acquisizione di unità terminologiche rilevanti per il dominio settoriale di acquisizione che siano distinte dalle unità che appartengono al lessico comune di una lingua. Tale difficoltà è dovuta principalmente alla stretta relazione che c'è tra lessico settoriale e lessico comune.

Affrontando la ben nota questione del rapporto biunivoco tra lingua comune e lingua settoriale, Beccaria in (Beccaria, 1973) individua in ciò che egli definisce «escursione terminologica» una caratteristica peculiare di ogni linguaggio settoriale. L'allusione è, da un lato, «alla crescente forza espansiva ed al prestigio reale nell'uso parlato e scritto di cui sono dotati i linguaggi settoriali», dall'altro, al fatto che «tra vocabolario comune e vocabolario tecnico si ergono sempre più esili barriere». Nonostante tali peculiarità condivise, egli sottolinea il fatto che «i modi e le direzioni dell'escursione terminologica sono peculiari del singolo linguaggio settoriale». Descrivendo infatti la natura del linguaggio giuridico «distinto ma non separato» dalla lingua comune, Mortara Garavelli in (Mortara Garavelli, 2001) ricorda che una cosa «è la condizione condivisa dalle varietà di lingua che differiscono dalla matrice comune per l'impiego di tecnicismi lessicali e per una formalità di registri», altra cosa è la «formalizzazione delle lingue speciali scientifiche».

2. La metodologia di estrazione

La metodologia di estrazione qui proposta, utilizzando strumenti di TAL e componenti di elaborazione statistica, mira ad acquisire da una collezione di testi di dominio le unità terminologiche settoriali. Inoltre, l'obiettivo di differenziare il lessico settoriale da quello comune è perseguito grazie all'approccio contrastivo seguito. Tale approccio si inserisce in un attivo filone di ricerca in materia di estrazione terminologica basato sul confronto della distribuzione di termini in corpora di dominio specialistico e in corpora di lingua comune (cfr. Basili et al., 2001).

Il processo di estrazione seguito si struttura nella seguente serie di passaggi consecutivi:

a) dato un testo in input, grazie ad una batteria di strumenti TAL, esso viene linguisticamente annotato rispetto a più livelli di analisi linguistica, i.e. segmentazione in unità-parola (*tokenization*), analisi morfo-sintattica (realizzata con l'analizzatore morfo-sintattico descritto in Dell'Orletta, 2009) e lemmatizzazione;
b) una lista di unità terminologiche monorematiche (es. *presidente*) e polirematiche (es. *presidente della repubblica*) candidate all'estrazione è definita a partire dal testo analizzato morfo-sintatticamente sulla base di una serie di filtri linguistici. In particolare, i) sono considerate come potenziali monorematiche le unità terminologiche annotate con la categoria morfo-sintattica 'nome' mentre ii) le unità polirematiche sono individuate sulla base di una serie di sequenze di categorie morfo-sintattiche rappresentative di diversi tipi di modificazione nominale;
c) ad ogni termine candidato viene associato un valore di significatività stabilito sulla base di filtri statistici. In particolare, per l'estrazione dei termini complessi è

applicato il filtro statistico C-NC Value (Frantzi et al., 1999), una delle misure più utilizzate in letteratura per determinare la probabilità di un'unità poliematica di essere un termine. La risultante lista di termini è ordinata per valori decrescenti;

d) una selezione dei primi termini di tale lista ordinata è confrontata con le occorrenze delle medesime unità terminologiche nel corpus di lingua comune usato come corpus di contrasto. Questo passaggio permette di riorganizzare i termini candidati all'estrazione rispetto ad un valore di contrasto calcolato sulla base di una funzione statistica. In questo modo il glossario terminologico finale contiene termini riorganizzati sulla base del valore di contrasto associato rispetto al corpus di lingua comune e non più sulla significatività statistica all'interno del corpus di dominio. Di conseguenza, i termini più significativi per il dominio settoriale avranno un valore di contrasto maggiore, mentre quelli meno significativi o appartenenti al lessico comune avranno valori più bassi.

3. I risultati di due esperimenti di estrazione

La metodologia descritta è stata sperimentata su due diversi domini settoriali, quello di storia dell'arte e quello giuridico, con risultati incoraggianti. In entrambi i casi, l'analisi dei glossari terminologici estratti ha dimostrato che il metodo contrastivo qui proposto è affidabile per estrarre termini caratterizzanti il dominio settoriale in questione, tenendoli distinti da quelli della lingua comune. Nel caso, ad esempio, del dominio artistico la fase di riordino dei termini candidati all'estrazione sulla base del valore di contrasto calcolato ha permesso di tenere distinte le unità terminologiche settoriali (es. *percorso espositivo, artista*, ecc…) da quelle della lingua comune (es. *istanza politica, definizione*, ecc…).

Inoltre, l'esperimento di estrazione condotto su un dominio settoriale profondamente diverso come quello giuridico ha messo in luce una seconda potenzialità della metodologia proposta: la possibilità cioè di discernere in casi di terminologia eterogenea, caratterizzata, come quella giuridica, da «tre polarità linguistiche: la lingua comune, utilizzata dalla totalità dei parlanti, la lingua speciale del diritto e altre lingue speciali, dipendenti da settori di conoscenze o sfere di attività specialistici». (cfr. Zaccaria, 2003). In questo caso, il metodo proposto ha permesso in una prima fase di isolare la terminologia comune (es. *direttore generale, anno seguente,* ecc…) e in una seconda fase distinguere il lessico giuridico (es. *decreto legislativo, norma*, ecc…) da quello del dominio regolato (es. *inquinamento atmosferico, effetto serra*, ecc…).

4. Bibliografia

Basili R., Moschetti A., Pazienza M.T. e Zanzotto FM. (2001), *A contrastive approach to term extraction*. In Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA–2001), Nancy.

Beccaria G.L. (1973), *Linguaggi settoriali e lingua comune*, in Beccaria G.L. (a cura di), *I linguaggi settoriali in Italia*, Milano, Bompiani, pp. 7-59.

Dell'Orletta F. (2009), *Ensemble system for Part-of-Speech tagging*, in Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, December.

Frantzi K. e Ananiadou S. (1999), *The C–value / NC Value domain independent method for multi-word term extraction*. In Journal of Natural Language Processing, 6(3):145–179.

Mortara Garavelli B. (2001), *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Einaudi, Torino.

Zaccaria G. (2003), *Testo, contesto e linguaggi settoriali nell'interpretazione giuridica*, in Mariani Marini A., *La lingua, la legge, la professione forense*, atti del convegno Accademia della Crusca (Firenze, 31 gennaio-1 febbraio 2002) Giuffrè, pp. 89-102.